

Low Rank Estimation of Similarities on Graphs

Vladimir Koltchinskii^{*} and Pedro Rangel[†]

School of Mathematics

Georgia Institute of Technology

Atlanta, GA 30332-0160

vlad@math.gatech.edu, prangel@math.gatech.edu

March 3, 2013

Abstract

Let (V, E) be a graph with vertex set V and edge set E . Let $(X, X', Y) \in V \times V \times \{-1, 1\}$ be a random triple, where X, X' are independent uniformly distributed vertices and Y is a label indicating whether X, X' are “similar” ($Y = +1$), or not ($Y = -1$). Our goal is to estimate the regression function

$$S_*(u, v) = \mathbb{E}(Y|X = u, X' = v), u, v \in V$$

based on training data consisting of n i.i.d. copies of (X, X', Y) . We are interested in this problem in the case when S_* is a symmetric low rank kernel and, in addition to this, it is assumed that S_* is “smooth” on the graph. We study estimators based on a modified least squares method with complexity penalization involving both the nuclear norm and Sobolev type norms of symmetric kernels on the graph and prove upper bounds on L_2 -type errors of such estimators with explicit dependence both on the rank of S_* and on the degree of its smoothness.

1 Introduction

Let $G = (V, E)$ be a graph with vertex set V and edge set E , $\text{card}(V) = m$. Let $A := (a(u, v))_{u, v \in V}$ be the adjacency matrix of G , that is, $a(u, v) = 1$ if u and v are connected with an edge and $a(u, v) = 0$ otherwise. Let $\Delta := D - A$ be the Laplacian of G , D being the diagonal matrix with the degrees of vertices on the diagonal. Let $(X, X', Y) \in V \times V \times \{-1, 1\}$ be a random triple with X, X' being independent vertices

^{*}Partially supported by NSF Grants DMS-1207808, DMS-0906880 and CCF-0808863

[†]Supported by NSF Grant CCF-0808863

sampled at random from the uniform distribution Π on V and Y being an “indicator” of a symmetric binary relationship between X, X' called in what follows a “similarity”. More precisely, $Y = +1$ indicates that the vertices X, X' are similar and $Y = -1$ indicates that they are not. The conditional distribution of Y given X, X' is completely characterized by the regression function

$$S_*(u, v) := \mathbb{E}(Y|X = u, X' = v), u, v \in V$$

that is assumed to be a symmetric kernel on $V \times V$ and will be called the *similarity kernel*. It is well known that $\text{sign}(S_*(X, X'))$ is the Bayes classifier, that is, the best possible predictor of Y based on an observation of X, X' in the sense that it minimizes the generalization error $\mathbb{P}\{Y \neq g(X, X')\}$ over all possible predictors $g : V \times V \mapsto \{-1, 1\}$. Our goal is to estimate S_* based on the training data $(X_1, X'_1, Y_1), \dots, (X_n, X'_n, Y_n)$ consisting of n i.i.d. copies of (X, X', Y) . We are especially interested in the class of problems such that, on the one hand, S_* is a matrix (kernel) of relatively small rank and, on the other hand, S_* possesses certain degree of smoothness on the graph.

Throughout the paper, \mathcal{S}_V denotes the linear space of *symmetric kernels* $S : V \times V \mapsto \mathbb{R}$, $S(u, v) = S(v, u)$, $u, v \in V$, that can be also viewed as real-valued symmetric $m \times m$ matrices. For $S \in \mathcal{S}_V$, let $\text{rank}(S)$ denote the *rank* of S and $\text{tr}(S)$ denote the *trace* of S . The *spectral representation* of S has the form $S = \sum_{j=1}^r \sigma_j (\psi_j \otimes \psi_j)$, where $r = \text{rank}(S)$, $\sigma_1 \leq \dots \leq \sigma_r$ are non-zero eigenvalues of S (repeated with their multiplicities) and ψ_1, \dots, ψ_r are the corresponding orthonormal eigenfunctions (there is a multiple choice of ψ_j s in the case of repeated eigenvalues). We also use the notation $\text{sign}(S) := \sum_{j=1}^r \text{sign}(\sigma_j) (\psi_j \otimes \psi_j)$ and we define the *support* of S , denoted by $\text{supp}(S)$, as the linear span of $\{\psi_1, \dots, \psi_r\}$ in \mathbb{R}^V .

For $1 \leq p < \infty$, the Schatten p -norm of $S \in \mathcal{S}_V$ is defined as

$$\|S\|_p := (\text{tr}(|S|^p))^{1/p} = \left(\sum_{j=1}^r |\sigma_j|^p \right)^{1/p},$$

where $|S| := \sqrt{S^2}$. For $p = 1$, $\|\cdot\|_1$ is called the *nuclear norm*, while, for $p = 2$, $\|\cdot\|_2$ is the *Hilbert–Schmidt* or *Frobenius norm*, that is, the norm induced by the Hilbert–Schmidt inner product which will be denoted by $\langle \cdot, \cdot \rangle$. The *operator* or *spectral norm* is defined as $\|S\| := \max_j |\sigma_j|$.

Let us also denote by $\Pi^2 := \Pi \times \Pi$ the distribution of random couple (X, X') in

$V \times V$ and let $\|S\|_{L_2(\Pi^2)}$ be the $L_2(\Pi^2)$ -norm of kernel S :

$$\|S\|_{L_2(\Pi^2)}^2 = \int_{V \times V} |S(u, v)|^2 \Pi^2(du, dv) = \mathbb{E}|S(X, X')|^2.$$

The corresponding inner product is denoted by $\langle \cdot, \cdot \rangle_{L_2(\Pi^2)}$. Clearly, under the assumption that the distribution Π is uniform in V , we have $\|S\|_{L_2(\Pi^2)} = m^{-2}\|S\|_2^2$ and $\langle S_1, S_2 \rangle_{L_2(\Pi^2)} = m^{-2}\langle S_1, S_2 \rangle$.

The smoothness of a symmetric kernel $S : V \times V \mapsto \mathbb{R}$ can be characterized in terms of Sobolev type norms $\|\Delta^{p/2}S\|_2^2$ for some $p > 0$. Note that if S is a kernel of rank r with spectral representation $S = \sum_{k=1}^r \mu_k(\psi_k \otimes \psi_k)$, then ¹

$$\|\Delta^{p/2}S\|_2^2 = \text{tr}(\Delta^{p/2}S^2\Delta^{p/2}) = \text{tr}(\Delta^p S^2) = \sum_{k=1}^m \mu_k^2 \langle \Delta^p \psi_k, \psi_k \rangle = \sum_{k=1}^m \mu_k^2 \|\Delta^{p/2} \psi_k\|^2,$$

so, essentially, the smoothness of the kernel S depends on the smoothness of its eigenfunctions ψ_k on the graph. In particular, for $p = 1$, we have

$$\|\Delta^{1/2}S\|_2^2 = \sum_{k=1}^m \mu_k^2 \sum_{u \sim v} |\psi_k(u) - \psi_k(v)|^2,$$

where the sum is over the couples of vertices connected with an edge.

Given a kernel S , let $L_n(S)$ denote the following penalized empirical risk:

$$\begin{aligned} L_n(S) &:= \left[\|S\|_{L_2(\Pi^2)}^2 - \frac{2}{n} \sum_{j=1}^n Y_j S(X_j, X'_j) + \varepsilon \|S\|_1 + \bar{\varepsilon} \|W^{1/2}S\|_{L_2(\Pi^2)}^2 \right] \\ &= \left[\|S\|_{L_2(\Pi^2)}^2 - \frac{2}{n} \sum_{j=1}^n Y_j S(X_j, X'_j) + \varepsilon \|S\|_1 + \varepsilon_1 \|W^{1/2}S\|_2^2 \right] \end{aligned} \quad (1.1)$$

where $W = d\Delta^p$ for some constants $d > 0$ and $p > 0$, $\varepsilon, \bar{\varepsilon} > 0$ are regularization parameters and $\varepsilon_1 = \frac{\bar{\varepsilon}}{m^2}$. We will study the following estimation method:

$$\hat{S} := \text{argmin}_{S \in \mathbb{D}} L_n(S), \quad (1.2)$$

where \mathbb{D} is a closed convex subset of the linear space \mathcal{S}_V of all symmetric kernels. Note that there are two complexity penalties involved in the definition of penalized empirical risk (1.1). The first penalty is based on the nuclear norm $\|S\|_1$ and it is used to “promote” low rank solutions. The second penalty is based on a “Sobolev type norm” $\|W^{1/2}S\|_2^2$.

¹Below $\|\cdot\|$ denotes the Euclidean norm in \mathbb{R}^V ; there is a little abuse of notation here since we also denote the operator norm by $\|\cdot\|$.

It is used to “promote” the smoothness of the solution on the graph. In principle, W in the definition of $L_n(S)$ could be an arbitrary symmetric nonnegatively definite matrix. Therefore, alternative interpretations of the problem under consideration are possible (such as, for instance, learning similarities on weighted graphs).

We will derive an upper bound on the error $\|\hat{S} - S_*\|_{L_2(\Pi^2)}^2 = m^{-2}\|\hat{S} - S_*\|_2^2$ of estimator \hat{S} in terms of spectral characteristics of the target similarity matrix S_* and matrix W . Before stating the main results, let us recall recent advances on low rank matrix completion problems in which the approach based on nuclear norm penalization has been crucial.

Suppose first that a symmetric kernel $S_* \in \mathcal{S}_V$ is observed at random points $(X_j, X'_j), j = 1, \dots, n$, where $X_j, X'_j, j = 1, \dots, n$ are independent and sampled from the uniform distribution Π in V . In this case, V is an arbitrary finite set of cardinality m and the set of edges E is not specified. It is assumed that $Y_j = S_*(X_j, X'_j)$, so, there is no errors in the observations. In such a noiseless case, the following method is used to recover S_* based on the observations $(X_1, X'_1, Y_1), \dots, (X_n, X'_n, Y_n)$:

$$\check{S} := \operatorname{argmin}\{\|S\|_1 : S \in \mathcal{S}_V, S(X_j, X'_j) = Y_j, j = 1, \dots, n\}.$$

Such methods of recovery of low rank target matrices S_* have been extensively studied in the recent literature (see Candes and Recht (2009), Recht, Fazel and Parrilo (2010), Candes and Tao (2010), Gross (2011) and references therein). It is easy to see that there are low rank matrices S_* that can not be recovered based on a random sample of n entries unless n is very large (comparable with the total number of entries of the matrix). Indeed, consider S_* such that, for given $u, v \in V$, $S_*(u, v) = S_*(v, u) = 1$ and $S_*(u', v') = 0$ otherwise. For this rank 2 matrix, the probability that the two “informative” entries are not present in the sample is $(1 - \frac{2}{m^2})^n$, which is close to 1 if $n = o(m^2)$. Such sparse low rank matrices should be excluded to make it possible to recover the target low rank matrix based on relatively small samples of entries. This is done by introducing so called *low coherence* assumptions. Let $\{e_v : v \in V\}$ be the canonical orthonormal basis of \mathbb{R}^V equipped with the standard Euclidean inner product. Given a linear subspace $L \subset \mathbb{R}^V$, denote by L^\perp the orthogonal complement of L and by P_L the projector onto the subspace L . Let $L := \operatorname{supp}(S_*)$, $r = \operatorname{rank}(S_*)$ and suppose there exists a constant $\nu \geq 1$ (*coherence coefficient*) such that

$$\|P_L e_v\|^2 \leq \frac{\nu r}{m}, \quad v \in V \quad \text{and} \quad |\langle \operatorname{sign}(S_*) e_u, e_v \rangle|^2 \leq \frac{\nu r}{m^2}, \quad u, v \in V. \quad (1.3)$$

The following result is due to Candes and Tao (2010) and Gross (2011) (we state

here a version of Gross that is an improvement of an earlier result of Candes and Tao with significant simplification of the proof).

Theorem 1 *Suppose conditions (1.3) hold for some $\nu \geq 1$. Then, there exists a constant $C > 0$ such that, for all $n \geq C\nu rm \log^2 m$, $\check{S} = S_*$ with probability at least $1 - m^{-2}$.*

Thus, if, for the target matrix S_* , the coherence coefficient $\nu \geq 1$ is relatively small, the nuclear norm minimization algorithm (1.2) does provide the exact recovery of S_* as soon as the number of observed entries n is of the order mr (up to a log factor).

In the case when Y_j are noisy observations of $S_*(X_j, X'_j)$ with

$$\mathbb{E}(Y_j | X_j = u, X'_j = v) = S_*(u, v),$$

one can use the following estimation method based on penalized empirical risk minimization with quadratic loss and with nuclear norm penalty:

$$\check{S} := \operatorname{argmin}_{S \in \mathcal{S}_V} \left[n^{-1} \sum_{j=1}^n (Y_j - S(X_j, X'_j))^2 + \varepsilon \|S\|_1 \right]. \quad (1.4)$$

This method has been also extensively studied for the recent years, in particular, by Candes and Plan (2011), Rohde and Tsybakov (2011), Negahban and Wainwright (2010), Koltchinskii, Lounici and Tsybakov (2011), Koltchinskii (2011b). It was also pointed out by Koltchinskii, Lounici and Tsybakov (2011) that in the case of known design distribution Π (which is the case in our paper) one can use instead of (1.4) the following modified method:²

$$\check{S} := \operatorname{argmin}_{S \in \mathcal{S}_V} \left[\|S\|_{L_2(\Pi^2)}^2 - \frac{2}{n} \sum_{j=1}^n Y_j S(X_j, X'_j) + \varepsilon \|S\|_1 \right]. \quad (1.5)$$

Clearly, (1.5) is equivalent to method (1.2) defined above for $\bar{\varepsilon} = 0$.

When the observations $|Y_j| \leq 1, j = 1, \dots, n$ (for instance, when $Y_j \in \{-1, 1\}$, which is the case studied in the paper), the next result follows from Theorem 4 in Koltchinskii, Lounici and Tsybakov (2011).

Theorem 2 *For $t > 0$, suppose that*

$$\varepsilon \geq 4 \left(\sqrt{\frac{t + \log(2m)}{nm}} \vee \frac{2(t + \log(2m))}{n} \right).$$

²Note that, if the norm $\|S\|_{L_2(\Pi^2)}$ in the definition below is replaced by the $L_2(\Pi_n)$ -norm, where Π_n is the empirical distribution based on $(X_1, X'_1), \dots, (X_n, X'_n)$, then the resulting estimator coincides with (1.4).

Then with probability at least $1 - e^{-t}$

$$\|\check{S} - S_*\|_{L_2(\Pi)}^2 \leq \left(\frac{1+\sqrt{2}}{2}\right)^2 m^2 \varepsilon^2 \text{rank}(S_*).$$

Our main goal is to show that this bound can be improved in the case when the target kernel S_* , in addition to having relatively small rank, is also smooth on the graph and when the estimation method (1.2) is used with a proper choice of regularization parameters $\varepsilon, \bar{\varepsilon}$.

2 Main Results

Suppose that W has the following spectral representation: $W = \sum_{k=1}^m \lambda_k (\phi_k \otimes \phi_k)$, where $0 \leq \lambda_1 \leq \dots \leq \lambda_m$ are the eigenvalues of W (repeated with their multiplicities) and ϕ_1, \dots, ϕ_m are the corresponding orthonormal eigenfunctions (of course, there is a multiple choice of ϕ_k in the case of repeated eigenvalues). Let k_0 be the smallest k such that $\lambda_k > 0$. We will assume that for some (arbitrarily large) $\zeta \geq 1$ $\lambda_m \leq m^\zeta$ and $\lambda_{k_0} \geq m^{-\zeta}$. In addition, it is assumed that $s \mapsto \frac{s}{\lambda_s}$ is a nonincreasing sequence, that, for all $k = k_0, \dots, m-1$, $\lambda_{k+1} \leq c\lambda_k$, and, that, for all $s \geq k_0$,

$$\sum_{k=s}^m \frac{1}{\lambda_k} \leq c \frac{s}{\lambda_s} \quad (2.1)$$

with a constant $c > 0$.

Suppose now that the spectral representation of S_* is $S_* = \sum_{k=1}^r \mu_k (\psi_k \otimes \psi_k)$, where $r = \text{rank}(S_*) \geq 1$, μ_k are non-zero eigenvalues of S_* (possibly repeated) and ψ_k are the corresponding orthonormal eigenfunctions. Denote $L := \text{supp}(S_*)$. Let φ be an arbitrary nondecreasing function such that $k \mapsto \frac{\varphi(k)}{k}$ is nonincreasing and

$$\sum_{j=1}^k \|P_L \phi_j\|^2 \leq \varphi(k), k = 0, 1, \dots, m.$$

We will denote by $\Psi = \Psi_{S_*, W}$ the class of all the functions satisfying these properties. Often, it will be convenient to extend a function $\varphi \in \Psi$ to nonnegative real numbers by making it linear in each of the intervals $[k, k+1], k = 0, 1, \dots, m-1$ and setting $\varphi(u) = \varphi(m)$ for all $u > m$. Such an extension will be also denoted by φ . It is easy to see that the extension is a nondecreasing function in \mathbb{R}_+ and the function $u \mapsto \frac{\varphi(u)}{u}$ is nonincreasing.

The following *coherence function* will be crucial in our analysis:

$$\bar{\varphi}(k) := \bar{\varphi}(S_*, k) := \max_{t \leq k} t \max_{j \geq t} \frac{1}{j} \sum_{i=1}^j \|P_L \phi_i\|^2, k = 1, \dots, m, \quad \bar{\varphi}(0) = 0.$$

It is straightforward to check that $\bar{\varphi} \in \Psi$ and, for all $\varphi \in \Psi$, $\bar{\varphi}(k) \leq \varphi(k)$, $k = 0, \dots, m$. Thus, $\bar{\varphi}$ is the smallest function $\varphi \in \Psi$. Also, $\bar{\varphi}(m) = r$ since $\sum_{j=1}^m \|P_L \phi_j\|^2 = \|P_L\|_2^2 = r$. Moreover, since $\frac{\bar{\varphi}(k)}{k}$ is nonincreasing, we have

$$\bar{\varphi}(k) \geq \frac{rk}{m}, k = 0, \dots, m.$$

Given $t > 0$, let $t_{n,m} := t + \log(2m(\log_2(4n^\zeta m^{(3/2)\zeta}) + 2))$. We will assume in what follows that $mt_{n,m} \leq n$ and set

$$\varepsilon := 4\sqrt{\frac{t + \log(2m)}{nm}}.$$

Theorem 3 *There exists constants C, C_1 depending only on c such that, for all $s \in \{k_0 + 1, \dots, m + 1\}$ and all $\bar{\varepsilon} \in [\lambda_s^{-1}, \lambda_{s-1}^{-1}]$,³ with probability at least $1 - e^{-t}$,*

$$\|\hat{S} - S_*\|_{L_2(\Pi^2)}^2 \leq C \frac{\bar{\varphi}(S_*; s) mt_{n,m}}{n} + \bar{\varepsilon} \|W^{1/2} S_*\|_{L_2(\Pi^2)}^2 + C_1 \max_{v \in V} \|P_L e_v\|^2 \left(\frac{mt_{n,m}}{n} \right)^2. \quad (2.2)$$

Remarks. Note that $\max_{v \in V} \|P_L e_v\|^2 \leq 1$. Thus, the last term in the righthand side of bound (2.2) is smaller than the first term, provided that

$$\bar{\varphi}(S_*; s) \geq \frac{mt_{n,m}}{n}.$$

Moreover, this term is much smaller under a low coherence condition $\max_{v \in V} \|P_L e_v\|^2 \leq \frac{\nu r}{m}$ for some $\nu \geq 1$ (see conditions (1.3)). In this case,

$$\max_{v \in V} \|P_L e_v\|^2 \left(\frac{mt_{n,m}}{n} \right)^2 \leq \frac{\nu r m t_{n,m}^2}{n^2} \leq \frac{\nu r t_{n,m}}{n}.$$

Note also that Theorem 3 holds in the case when $\bar{\varepsilon} = 0$. In this case, $s = m$ and $\bar{\varphi}(S_*, m) = r$, so the bound of Theorem 3 becomes

$$\|\hat{S} - S_*\|_{L_2(\Pi^2)}^2 \leq C \frac{r m t_{n,m}}{n}, \quad (2.3)$$

which also follows from the result of Koltchinskii, Lounici and Tsybakov (2011) (see Theorem 2 in Section 1).

³Here and in what follows, we use a convention that $\lambda_{m+1} = +\infty$ and $\lambda_{m+1}^{-1} = 0$.

The function $\bar{\varphi}$ involved in the statement of the theorem has some connection to the low coherence assumptions frequently used in the literature on low rank matrix completion. To be specific, suppose that, for some $\nu \geq 1$,

$$\sum_{j=1}^k \|P_L \phi_j\|^2 \leq \frac{\nu r k}{m}, k = 1, \dots, m. \quad (2.4)$$

Then

$$\bar{\varphi}(k) \leq \frac{\nu r k}{m}, k = 1, \dots, m.$$

A part of standard low coherence assumptions on matrix S_* with respect to the orthonormal basis $\{\phi_k\}$ is (see (1.3))

$$\|P_L \phi_k\|^2 \leq \frac{\nu r}{m}, k = 1, \dots, m$$

and it implies condition (2.4) that can be viewed as a weak version of low coherence. Under condition (2.4), the following corollary of Theorem 3 holds.

Corollary 1 *Suppose that condition (2.4) holds. Then, there exists a constant $C > 0$ depending only on ζ such that, for all $s \in \{k_0 + 1, \dots, m + 1\}$ and all $\bar{\varepsilon} \in (\lambda_s^{-1}, \lambda_{s-1}^{-1}]$, with probability at least $1 - e^{-t}$,*

$$\|\hat{S} - S_*\|_{L_2(\Pi^2)}^2 \leq C \frac{\nu r s t_{n,m}}{n} + \bar{\varepsilon} \|W^{1/2} S_*\|_{L_2(\Pi^2)}^2 + C_1 \max_{v \in V} \|P_L e_v\|^2 \left(\frac{m t_{n,m}}{n} \right)^2.$$

Note that, if $\lambda_k \asymp k^{2\beta}$ for some $\beta > 1/2$, then the choice of s that minimizes the bound of Corollary 1 is $s \asymp \left(\frac{n}{\nu r t_{n,m}} \right)^{1/(2\beta+1)} \|W^{1/2} S_*\|_{L_2(\Pi)}^{2/(2\beta+1)}$, which, under a low coherence assumption $\max_{v \in V} \|P_L e_v\|^2 \leq \frac{\nu r}{m}$, yields the bound

$$\|\hat{S} - S_*\|_{L_2(\Pi^2)}^2 \leq C \left(\frac{\nu r t_{n,m}}{n} \right)^{2\beta/(2\beta+1)} \|W^{1/2} S_*\|_{L_2(\Pi)}^{2/(2\beta+1)}. \quad (2.5)$$

The advantage of (2.5) comparing with (2.3) (that holds for $\bar{\varepsilon} = 0$ and does not rely on any smoothness assumption on the kernel S_*) is due to the fact that there is no factor m in the numerator in the right hand side of (2.5). Due to this fact, when m is large enough and ν is not too large, bound (2.5) becomes sharper than (2.3).

3 Proofs

Proof of Theorem 3. Bound (2.2) will be proved for an arbitrary function $\varphi \in \Psi_{S_*, W}$ with $\varphi(k) = r, k \geq m$ instead of $\bar{\varphi}$. It then can be applied to the function $\bar{\varphi}$ (which

is the smallest function in $\Psi_{S_*, W}$). We will also assume throughout the proof that $s \in \{k_0, \dots, m\}$ and $\bar{\varepsilon} \in [\lambda_{s+1}^{-1}, \lambda_s^{-1}]$ (at the end of the proof, we replace $s+1 \mapsto s$).

Denote $\mathcal{P}_L(A) := A - P_{L^\perp} A P_{L^\perp}$, $\mathcal{P}_L^\perp(A) = P_{L^\perp} A P_{L^\perp}$, $A \in \mathcal{S}_V$. Clearly, this defines orthogonal projectors $\mathcal{P}_L, \mathcal{P}_L^\perp$ in the space \mathcal{S}_V with Hilbert–Schmidt inner product. We will use the following well known representation of subdifferential of convex function $S \mapsto \|S\|_1 : \partial\|S\|_1 = \{\text{sign}(S) + \mathcal{P}_L^\perp(M) : M \in \mathcal{S}_V, \|M\| \leq 1\}$, where $L = \text{supp}(S)$ (see Koltchinskii (2011b), Appendix A.4 and references therein). An arbitrary matrix $A \in \partial L_n(\hat{S})$ can be represented as follows:

$$A = \frac{2}{m^2} \hat{S} - \frac{2}{n} \sum_{i=1}^n Y_i E_{X_i, X'_i} + \varepsilon \hat{V} + 2\varepsilon_1 W \hat{S}, \quad (3.1)$$

where $\hat{V} \in \partial\|\hat{S}\|_1$ and $E_{u,v} = E_{v,u} = \frac{1}{2}(e_u \otimes e_v + e_v \otimes e_u)$. Since \hat{S} is a minimizer of $L_n(S)$, there exists a matrix $A \in \partial L_n(\hat{S})$ such that $-A$ belongs to the normal cone of \mathbb{D} at the point \hat{S} (see Aubin and Ekeland (1984), Chap. 2, Corollary 6). This implies that $\langle A, \hat{S} - S_* \rangle \leq 0$ and, in view of (3.1),

$$2\langle \hat{S}, \hat{S} - S_* \rangle_{L_2(\Pi^2)} - \left\langle \frac{2}{n} \sum_{i=1}^n Y_i E_{X_i, X'_i}, \hat{S} - S_* \right\rangle + \varepsilon \langle \hat{V}, \hat{S} - S_* \rangle + 2\varepsilon_1 \langle W \hat{S}, \hat{S} - S_* \rangle \leq 0$$

It follows by a simple algebra that

$$\begin{aligned} & 2\|\hat{S} - S_*\|_{L_2(\Pi^2)}^2 + 2\varepsilon_1 \|W^{1/2}(\hat{S} - S_*)\|_2^2 + \varepsilon \langle \hat{V}, \hat{S} - S_* \rangle \\ & \leq -2\varepsilon_1 \langle S_*, W(\hat{S} - S_*) \rangle + 2\langle \Xi, \hat{S} - S_* \rangle, \end{aligned} \quad (3.2)$$

where

$$\Xi := \frac{1}{n} \sum_{j=1}^n Y_j E_{X_j, X'_j} - \mathbb{E} Y E_{X, X'}.$$

Note that $\langle \Xi, S \rangle = \frac{1}{n} \sum_{j=1}^n (Y_j S(X_j, X'_j) - \mathbb{E} Y S(X, X'))$.

On the other hand, let $V_* \in \partial\|S_*\|_1$. Therefore, the representation $V_* = \text{sign}(S_*) + \mathcal{P}_L^\perp(M)$ holds, where M is a matrix with $\|M\| \leq 1$. It follows from the trace duality property that there exists an M with $\|M\| \leq 1$ such that

$$\langle \mathcal{P}_L^\perp(M), \hat{S} - S_* \rangle = \langle M, \mathcal{P}_L^\perp(\hat{S} - S_*) \rangle = \langle M, \mathcal{P}_L^\perp(\hat{S}) \rangle = \|\mathcal{P}_L^\perp(\hat{S})\|_1$$

where in the first equality we used that \mathcal{P}_L^\perp is a self-adjoint operator and in the second equality we used that S_* has support L . Using this equation and monotonicity of subdifferentials of convex functions, we get

$$\langle \text{sign}(S_*), \hat{S} - S_* \rangle + \|\mathcal{P}_L^\perp(\hat{S})\|_1 = \langle V_*, \hat{S} - S_* \rangle \leq \langle \hat{V}, \hat{S} - S_* \rangle$$

Substituting this in (3.2), it is easy to get

$$\begin{aligned} & 2\|\hat{S} - S_*\|_{L_2(\Pi^2)}^2 + \varepsilon\|\mathcal{P}_L^\perp(\hat{S})\|_1 + 2\varepsilon_1\|W^{1/2}(\hat{S} - S_*)\|_2^2 \leq \\ & -\varepsilon\langle \text{sign}(S_*), \hat{S} - S_* \rangle - 2\varepsilon_1\langle W^{1/2}S_*, W^{1/2}(\hat{S} - S_*) \rangle + 2\langle \Xi, \hat{S} - S_* \rangle \end{aligned} \quad (3.3)$$

We will bound separately each term in the right hand side. First note that

$$\begin{aligned} & \varepsilon|\langle \text{sign}(S_*), \hat{S} - S_* \rangle| \leq \varepsilon\|\text{sign}(S_*)\|_2\|\hat{S} - S_*\|_2 \\ & = \varepsilon\sqrt{r}m\|\hat{S} - S_*\|_{L_2(\Pi^2)} \leq \frac{1}{2}rm^2\varepsilon^2 + \frac{1}{2}\|\hat{S} - S_*\|_{L_2(\Pi^2)}^2. \end{aligned} \quad (3.4)$$

We will also need a more subtle bound on $\langle \text{sign}(S_*), \hat{S} - S_* \rangle$, expressed in terms of function φ . Note that, for all $k_0 \leq s \leq m$,

$$\begin{aligned} \langle \text{sign}(S_*), \hat{S} - S_* \rangle &= \sum_{k=1}^m \langle \text{sign}(S_*)\phi_k, (\hat{S} - S_*)\phi_k \rangle = \\ &= \sum_{k=1}^s \langle \text{sign}(S_*)\phi_k, (\hat{S} - S_*)\phi_k \rangle + \sum_{k=s+1}^m \left\langle \frac{\text{sign}(S_*)\phi_k}{\sqrt{\lambda_k}}, \sqrt{\lambda_k}(\hat{S} - S_*)\phi_k \right\rangle, \end{aligned}$$

which easily implies

$$\begin{aligned} |\langle \text{sign}(S_*), \hat{S} - S_* \rangle| &\leq \left(\sum_{k=1}^s \|\text{sign}(S_*)\phi_k\|^2 \right)^{1/2} \left(\sum_{k=1}^s \|(\hat{S} - S_*)\phi_k\|^2 \right)^{1/2} + \\ &\left(\sum_{k=s+1}^m \frac{\|\text{sign}(S_*)\phi_k\|^2}{\lambda_k} \right)^{1/2} \left(\sum_{k=s+1}^m \lambda_k \|(\hat{S} - S_*)\phi_k\|^2 \right)^{1/2} \leq \\ &\left(\sum_{k=1}^s \|P_L\phi_k\|^2 \right)^{1/2} \|\hat{S} - S_*\|_2 + \left(\sum_{k=s+1}^m \frac{\|P_L\phi_k\|^2}{\lambda_k} \right)^{1/2} \|W^{1/2}(\hat{S} - S_*)\|_2. \end{aligned} \quad (3.5)$$

We will now use the following elementary lemma.

Lemma 1 *Let c be the constant from condition (2.1). For all $s \geq k_0 - 1$,*

$$\sum_{k=s+1}^m \frac{\|P_L\phi_k\|^2}{\lambda_k} \leq (c+2) \frac{\varphi(s+1)}{\lambda_{s+1}}.$$

Proof. Denote $F_s := \sum_{k=1}^s \|P_L\phi_k\|^2$, $s = 1, \dots, m$. Then, using the properties of

function $\varphi \in \Psi$, we get

$$\begin{aligned} \sum_{k=s+1}^m \frac{\|P_L \phi_k\|^2}{\lambda_k} &= \sum_{k=s+1}^{m-1} F_k \left(\frac{1}{\lambda_k} - \frac{1}{\lambda_{k+1}} \right) + \frac{F_m}{\lambda_m} - \frac{F_s}{\lambda_{s+1}} \leq \\ &\sum_{k=s+1}^{m-1} \varphi(k) \left(\frac{1}{\lambda_k} - \frac{1}{\lambda_{k+1}} \right) + \frac{\varphi(m)}{\lambda_m} \leq \frac{\varphi(s+1)}{s+1} \left[\sum_{k=s+1}^{m-1} k \left(\frac{1}{\lambda_k} - \frac{1}{\lambda_{k+1}} \right) + \frac{m}{\lambda_m} \right] \leq \\ &\frac{\varphi(s+1)}{s+1} \left[\sum_{k=s+2}^m \frac{k - (k-1)}{\lambda_k} + \frac{(s+1)}{\lambda_{s+1}} + \frac{m}{\lambda_m} \right] = \frac{\varphi(s+1)}{s+1} \left[\sum_{k=s+2}^m \frac{1}{\lambda_k} + \frac{(s+1)}{\lambda_{s+1}} + \frac{m}{\lambda_m} \right]. \end{aligned}$$

Using the assumptions on the spectrum of W (in particular, condition (2.1)), we conclude that

$$\sum_{k=s+1}^m \frac{\|P_L \phi_k\|^2}{\lambda_k} \leq \frac{\varphi(s+1)}{s+1} \left[c \frac{s+1}{\lambda_{s+1}} + \frac{(s+1)}{\lambda_{s+1}} + \frac{m}{\lambda_m} \right] \leq (c+2) \frac{\varphi(s+1)}{\lambda_{s+1}},$$

ending the proof. \square

It follows from (3.5) and the bound of Lemma 1 that

$$\begin{aligned} |\langle \text{sign}(S_*), \hat{S} - S_* \rangle| &\leq \sqrt{\varphi(s)} \|\hat{S} - S_*\|_2 + \sqrt{(c+2) \frac{\varphi(s+1)}{\lambda_{s+1}}} \|W^{1/2}(\hat{S} - S_*)\|_2 = \\ &m \sqrt{\varphi(s)} \|\hat{S} - S_*\|_{L_2(\Pi^2)} + m \sqrt{(c+2) \frac{\varphi(s+1)}{\lambda_{s+1}}} \|W^{1/2}(\hat{S} - S_*)\|_{L_2(\Pi^2)}. \end{aligned} \quad (3.6)$$

This implies the following bound:

$$\begin{aligned} \varepsilon |\langle \text{sign}(S_*), \hat{S} - S_* \rangle| &\leq \\ \varphi(s) m^2 \varepsilon^2 + \frac{1}{4} \|\hat{S} - S_*\|_{L_2(\Pi^2)}^2 &+ (c+2) \frac{\varphi(s+1)}{\lambda_{s+1}} \frac{m^2 \varepsilon^2}{\bar{\varepsilon}} + \frac{\bar{\varepsilon}}{4} \|W^{1/2}(\hat{S} - S_*)\|_{L_2(\Pi^2)}^2, \end{aligned} \quad (3.7)$$

where we used twice an elementary inequality $ab \leq a^2 + \frac{1}{4}b^2$, $a, b > 0$. Since, under the assumptions of the theorem, $\bar{\varepsilon} \lambda_{s+1} \geq 1$, (3.7) yields the following bound:

$$\begin{aligned} \varepsilon |\langle \text{sign}(S_*), \hat{S} - S_* \rangle| &\leq \\ (c+3) \varphi(s+1) m^2 \varepsilon^2 + \frac{1}{4} \|\hat{S} - S_*\|_{L_2(\Pi^2)}^2 &+ \frac{\bar{\varepsilon}}{4} \|W^{1/2}(\hat{S} - S_*)\|_{L_2(\Pi^2)}^2. \end{aligned} \quad (3.8)$$

To bound the second term in the right hand side of (3.3), note that

$$|\langle W^{1/2} S_*, W^{1/2}(\hat{S} - S_*) \rangle| \leq \|W^{1/2} S_*\|_2 \|W^{1/2}(\hat{S} - S_*)\|_2, \quad (3.9)$$

which implies

$$\begin{aligned} \varepsilon_1 |\langle W^{1/2} S_*, W^{1/2} (\hat{S} - S_*) \rangle| &\leq \varepsilon_1 \|W^{1/2} S_*\|_2^2 + \frac{\varepsilon_1}{4} \|W^{1/2} (\hat{S} - S_*)\|_2^2 = \\ \bar{\varepsilon} \|W^{1/2} S_*\|_{L_2(\Pi^2)}^2 + \frac{\bar{\varepsilon}}{4} \|W^{1/2} (\hat{S} - S_*)\|_{L_2(\Pi^2)}^2. \end{aligned} \quad (3.10)$$

Finally, we bound $\langle \Xi, \hat{S} - S_* \rangle$:

$$\begin{aligned} |\langle \Xi, \hat{S} - S_* \rangle| &\leq |\langle \Xi, \mathcal{P}_L(\hat{S} - S_*) \rangle| + |\langle \Xi, \mathcal{P}_L^\perp(\hat{S}) \rangle| \\ &\leq |\langle \mathcal{P}_L \Xi, \hat{S} - S_* \rangle| + \|\Xi\| \|\mathcal{P}_L^\perp(\hat{S})\|_1. \end{aligned} \quad (3.11)$$

To bound $\|\Xi\|$, we use a version of noncommutative Bernstein inequality of Ahlswede and Winter (2002) (see also Tropp (2010), Koltchinskii (2011a, 2011b, 2011c) for other versions of such inequalities).

Lemma 2 *Let Z be a bounded random symmetric matrix with $\mathbb{E}Z = 0$, $\sigma_Z^2 := \|\mathbb{E}Z^2\|$ and $\|Z\| \leq U$ for some $U > 0$. Let Z_1, \dots, Z_n be n i.i.d. copies of Z . Then for all $t > 0$, with probability at least $1 - e^{-t}$*

$$\left\| \frac{1}{n} \sum_{i=1}^n Z_i \right\| \leq 2 \left(\sigma_Z \sqrt{\frac{t + \log(2m)}{n}} \vee U \frac{t + \log(2m)}{n} \right)$$

It is applied to i.i.d. random matrices $Z_i := Y_i E_{X_i, X'_i} - \mathbb{E}(Y_i E_{X_i, X'_i})$, $i = 1, \dots, n$. Since $\|Z_i\| \leq 2$ and, by a simple computation, $\sigma_{Z_i}^2 := \|\mathbb{E}Z_i^2\| \leq 1/m$ (see, e.g., Koltchinskii (2011b), Section 9.4), Lemma 2 implies that with probability at least $1 - e^{-t}$

$$\|\Xi\| = \left\| \frac{1}{n} \sum_{i=1}^n Z_i \right\| \leq 2 \left[\sqrt{\frac{t + \log(2m)}{nm}} \vee \frac{2(t + \log(2m))}{n} \right].$$

Under the assumption that

$$\varepsilon \geq 4 \left[\sqrt{\frac{t + \log(2m)}{nm}} \vee \frac{2(t + \log(2m))}{n} \right],$$

this yields $\|\Xi\| \leq \varepsilon/2$ and

$$|\langle \Xi, \hat{S} - S_* \rangle| \leq |\langle \mathcal{P}_L \Xi, \hat{S} - S_* \rangle| + \frac{\varepsilon}{2} \|\mathcal{P}_L^\perp(\hat{S})\|_1. \quad (3.12)$$

For simplicity, it is assumed that $n \geq 2m(t + \log(2m))$. In this case, one can take $\varepsilon = 4 \sqrt{\frac{t + \log(2m)}{nm}}$, as it has been done in the statement of the theorem.

We have to bound $|\langle \mathcal{P}_L \Xi, \hat{S} - S_* \rangle|$ and we start with the following simple bound:

$$\begin{aligned}
|\langle \mathcal{P}_L \Xi, \hat{S} - S_* \rangle| &\leq m \|\mathcal{P}_L \Xi\|_2 \|\hat{S} - S_*\|_{L^2(\Pi^2)} \\
&\leq m \sqrt{2r} \|\Xi\| \|\hat{S} - S_*\|_{L^2(\Pi^2)} \\
&\leq \frac{1}{2} m \varepsilon \sqrt{2r} \|\hat{S} - S_*\|_{L^2(\Pi^2)} \\
&\leq \frac{1}{2} m^2 \varepsilon^2 r + \frac{1}{4} \|\hat{S} - S_*\|_{L^2(\Pi^2)}^2,
\end{aligned} \tag{3.13}$$

where we use the fact that $\text{rank}(\mathcal{P}_L \Xi) \leq 2r$. Substituting (3.4), (3.10), (3.12) and (3.13) in (3.3), we easily get that

$$\|\hat{S} - S_*\|_{L^2(\Pi^2)}^2 \leq \frac{3}{2} r \varepsilon^2 m^2 + 2\varepsilon \|W^{1/2} S_*\|_{L^2(\Pi^2)}^2. \tag{3.14}$$

For $\varepsilon = 0$, this bound follows from the results of Koltchinskii, Lounici and Tsybakov (2011). However, we need a more subtle bound expressed in terms of function φ , which is akin to bound (3.8). To this end, we will use the following lemma.

Lemma 3 *For $\delta > 0$, let $k(\delta)$ be the largest value of $k \leq m$ such that $\lambda_k^{-1} \geq \delta^2$ (if $\lambda_1^{-1} < \delta^2$, we set $k(\delta) = 0$). For all $t > 0$, with probability at least $1 - e^{-t}$,*

$$\sup_{\|M\|_2 \leq \delta, \|W^{1/2} M\|_2 \leq 1} |\langle \mathcal{P}_L \Xi, M \rangle| \leq 2\sqrt{(4c+8)} \sqrt{\frac{t}{nm}} \delta \sqrt{\varphi(k(\delta)+1)} + 2\sqrt{2} \delta \max_{v \in V} \|P_L e_v\| \frac{t}{n},$$

provided that $k(\delta) < m$, and

$$|\langle \mathcal{P}_L \Xi, M \rangle| \leq 4\sqrt{2} \delta \sqrt{\frac{rt}{nm}} + 2\sqrt{2} \delta \max_{v \in V} \|P_L e_v\| \frac{t}{n},$$

provided that $k(\delta) \geq m$.

Proof. The proof is somewhat akin to the derivation of the bounds on Rademacher processes in terms of Mendelson's complexities used in learning theory (see, e.g., Proposition 3.3 in Koltchinskii (2011b)).

Note that, for all symmetric $m \times m$ matrices M ,

$$\langle \mathcal{P}_L \Xi, M \rangle = \sum_{k,j=1}^m \langle \mathcal{P}_L \Xi, \phi_k \otimes \phi_j \rangle \langle M, \phi_k \otimes \phi_j \rangle.$$

Suppose that

$$\|M\|_2^2 = \sum_{k,j=1}^m |\langle M, \phi_k \otimes \phi_j \rangle|^2 \leq \delta^2$$

and

$$\|W^{1/2}M\|_2^2 = \sum_{k,j=1}^m \lambda_k |\langle M, \phi_k \otimes \phi_j \rangle|^2 \leq 1.$$

Then, it easily follows that

$$\sum_{k,j=1}^m \frac{|\langle M, \phi_k \otimes \phi_j \rangle|^2}{\lambda_k^{-1} \wedge \delta^2} \leq 2,$$

which implies

$$\begin{aligned} |\langle \mathcal{P}_L \Xi, M \rangle| &\leq \\ &\left(\sum_{k,j=1}^m (\lambda_k^{-1} \wedge \delta^2) |\langle \mathcal{P}_L \Xi, \phi_k \otimes \phi_j \rangle|^2 \right)^{1/2} \left(\sum_{k,j=1}^m \frac{|\langle M, \phi_k \otimes \phi_j \rangle|^2}{\lambda_k^{-1} \wedge \delta^2} \right)^{1/2} \leq \\ &\sqrt{2} \left(\sum_{k,j=1}^m (\lambda_k^{-1} \wedge \delta^2) |\langle \mathcal{P}_L \Xi, \phi_k \otimes \phi_j \rangle|^2 \right)^{1/2}. \end{aligned} \tag{3.15}$$

Define now the following inner product:

$$\langle M_1, M_2 \rangle_w := \sum_{k,j=1}^m (\lambda_k^{-1} \wedge \delta^2) \langle M_1, \phi_k \otimes \phi_j \rangle \langle M_2, \phi_k \otimes \phi_j \rangle$$

and let $\|\cdot\|_w$ be the corresponding norm. We will provide an upper bound on

$$\|\mathcal{P}_L \Xi\|_w = \left(\sum_{k,j=1}^m (\lambda_k^{-1} \wedge \delta^2) |\langle \mathcal{P}_L \Xi, \phi_k \otimes \phi_j \rangle|^2 \right)^{1/2}.$$

To this end, we use a standard Bernstein type inequality for random variables in a Hilbert space. It is given in the following lemma.

Lemma 4 *Let ξ be a bounded random variable with values in a Hilbert space \mathcal{H} . Suppose that $\mathbb{E}\xi = 0$, $\mathbb{E}\|\xi\|_{\mathcal{H}}^2 = \sigma^2$ and $\|\xi\|_{\mathcal{H}} \leq U$. Let ξ_1, \dots, ξ_n be n i.i.d. copies of ξ . Then for all $t > 0$, with probability at least $1 - e^{-t}$*

$$\left\| \frac{1}{n} \sum_{i=1}^n \xi_i \right\|_{\mathcal{H}} \leq 2 \left[\sigma \sqrt{\frac{t}{n}} \vee U \frac{t}{n} \right]$$

Applying Lemma 4 to the random variable $\xi = Y\mathcal{P}_L(E_{X,X'}) - \mathbb{E}Y\mathcal{P}_L(E_{X,X'})$, we get that for all $t > 0$, with probability at least $1 - e^{-t}$,

$$\begin{aligned} \|\mathcal{P}_L \Xi\|_w &= \left\| \frac{1}{n} \sum_{j=1}^n Y_j \mathcal{P}_L(E_{X_j, X'_j}) - \mathbb{E}Y\mathcal{P}_L(E_{X,X'}) \right\|_w \leq \\ &2 \left[\mathbb{E}^{1/2} \|Y\mathcal{P}_L(E_{X,X'})\|_w^2 \sqrt{\frac{t}{n}} + \left\| Y\mathcal{P}_L(E_{X,X'}) \right\|_w \left\| \frac{t}{n} \right\|_{L_\infty} \right]. \end{aligned} \tag{3.16}$$

Using the fact that $Y \in \{-1, 1\}$, we get

$$\begin{aligned}
& \mathbb{E} \|Y \mathcal{P}_L(E_{X, X'})\|_w^2 = \mathbb{E} \|\mathcal{P}_L(E_{X, X'})\|_w^2 = \\
& \mathbb{E} \sum_{k,j=1}^m (\lambda_k^{-1} \wedge \delta^2) |\langle \mathcal{P}_L(E_{X, X'}), \phi_k \otimes \phi_j \rangle|^2 = \sum_{k,j=1}^m (\lambda_k^{-1} \wedge \delta^2) \mathbb{E} |\langle E_{X, X'}, \mathcal{P}_L(\phi_k \otimes \phi_j) \rangle|^2 = \\
& \sum_{k,j=1}^m (\lambda_k^{-1} \wedge \delta^2) m^{-2} \sum_{u,v \in V} |\langle E_{u,v}, \mathcal{P}_L(\phi_k \otimes \phi_j) \rangle|^2 \leq \\
& m^{-2} \sum_{k,j=1}^m (\lambda_k^{-1} \wedge \delta^2) \|\mathcal{P}_L(\phi_k \otimes \phi_j)\|_2^2 \leq 2m^{-2} \sum_{k,j=1}^m (\lambda_k^{-1} \wedge \delta^2) (\|P_L \phi_k\|^2 + \|P_L \phi_j\|^2) = \\
& 2m^{-1} \sum_{k=1}^m (\lambda_k^{-1} \wedge \delta^2) \|P_L \phi_k\|^2 + 2m^{-2} \sum_{k=1}^m (\lambda_k^{-1} \wedge \delta^2) \sum_{j=1}^m \|P_L \phi_j\|^2 = \\
& 2m^{-1} \sum_{k=1}^m (\lambda_k^{-1} \wedge \delta^2) \|P_L \phi_k\|^2 + 2m^{-2} \sum_{k=1}^m (\lambda_k^{-1} \wedge \delta^2) \|P_L\|_2^2 = \\
& 2m^{-1} \sum_{k=1}^m (\lambda_k^{-1} \wedge \delta^2) \|P_L \phi_k\|^2 + 2m^{-2} r \sum_{k=1}^m (\lambda_k^{-1} \wedge \delta^2).
\end{aligned} \tag{3.17}$$

To bound $\mathbb{E} \|Y \mathcal{P}_L(E_{X, X'})\|_w^2$ further, note that

$$\sum_{k=1}^m (\lambda_k^{-1} \wedge \delta^2) \|P_L \phi_k\|^2 \leq \delta^2 \sum_{k \leq k(\delta)} \|P_L \phi_k\|^2 + \sum_{k > k(\delta)} \lambda_k^{-1} \|P_L \phi_k\|^2. \tag{3.18}$$

Assuming that $1 \leq k(\delta) \leq m-1$, using the bound of Lemma 1, the fact that $\lambda_{k(\delta)+1}^{-1} < \delta^2$ and the monotonicity of function φ , we get from (3.18) that

$$\begin{aligned}
& \sum_{k=1}^m (\lambda_k^{-1} \wedge \delta^2) \|P_L \phi_k\|^2 \leq \delta^2 \varphi(k(\delta)) + (c+2) \frac{\varphi(k(\delta)+1)}{\lambda_{k(\delta)+1}} \leq \\
& \delta^2 \varphi(k(\delta)) + (c+2) \delta^2 \varphi(k(\delta)+1) \leq (c+3) \delta^2 \varphi(k(\delta)+1).
\end{aligned} \tag{3.19}$$

It is easy to check that (3.19) holds also for $k(\delta) = 0$ and $k(\delta) = m$ (in the last case, $\varphi(k(\delta)+1) = r$). We also have

$$\sum_{k=1}^m (\lambda_k^{-1} \wedge \delta^2) \leq \sum_{k \leq k(\delta)} \delta^2 + \sum_{k > k(\delta)} \lambda_k^{-1},$$

which, in view of condition (2.1), implies

$$\sum_{k=1}^m (\lambda_k^{-1} \wedge \delta^2) \leq \delta^2 k(\delta) + c \frac{k(\delta)+1}{\lambda_{k(\delta)+1}} \leq (c+1) \delta^2 (k(\delta)+1). \tag{3.20}$$

Using bounds (3.17), (3.19) and (3.20), we get, under the condition that $k(\delta) < m$,

$$\begin{aligned}
\mathbb{E}\|Y\mathcal{P}_L(E_{X,X'})\|_w^2 &\leq \\
2m^{-1}(c+3)\delta^2\varphi(k(\delta)+1) + 2m^{-2}r(c+1)\delta^2(k(\delta)+1) &\leq \\
2m^{-1}(c+3)\delta^2\varphi(k(\delta)+1) + 2m^{-2}r(c+1)\delta^2\frac{k(\delta)+1}{\varphi(k(\delta)+1)}\varphi(k(\delta)+1) &\leq \\
2m^{-1}(c+3)\delta^2\varphi(k(\delta)+1) + 2m^{-2}r(c+1)\delta^2\frac{m}{\varphi(m)}\varphi(k(\delta)+1) &= \\
(4c+8)m^{-1}\delta^2\varphi(k(\delta)+1). &
\end{aligned} \tag{3.21}$$

In the case when $k(\delta) \geq m$, it is easy to show that

$$\mathbb{E}\|Y\mathcal{P}_L(E_{X,X'})\|_w^2 \leq 4m^{-1}\delta^2r. \tag{3.22}$$

We can also bound $\left\|\|Y\mathcal{P}_L(E_{X,X'})\|_w\right\|_{L_\infty}^2$ as follows:

$$\begin{aligned}
\left\|\|Y\mathcal{P}_L(E_{X,X'})\|_w\right\|_{L_\infty}^2 &= \left\|\|\mathcal{P}_L(E_{X,X'})\|_w\right\|_{L_\infty}^2 = \\
\left\|\sum_{k,j=1}^m (\lambda_k^{-1} \wedge \delta^2) |\langle \mathcal{P}_L(E_{X,X'}), \phi_k \otimes \phi_j \rangle|^2\right\|_{L_\infty} &\leq \\
\max_{1 \leq k \leq m} (\lambda_k^{-1} \wedge \delta^2) \max_{u,v \in V} \sum_{k,j=1}^m |\langle \mathcal{P}_L E_{u,v}, \phi_k \otimes \phi_j \rangle|^2 &\leq \\
\max_{1 \leq k \leq m} (\lambda_k^{-1} \wedge \delta^2) \max_{u,v \in V} \|\mathcal{P}_L E_{u,v}\|_2^2 &\leq \delta^2 \max_{u,v \in V} \|\mathcal{P}_L(e_u \otimes e_v)\|_2^2 \leq 2\delta^2 \max_{v \in V} \|P_L e_v\|^2.
\end{aligned} \tag{3.23}$$

If $k(\delta) < m$, it follows from (3.15), (3.16), (3.21) and (3.23) that with probability at least $1 - e^{-t}$, for all symmetric matrices M with $\|M\|_2 \leq \delta$ and $\|W^{1/2}M\|_2 \leq 1$,

$$|\langle \mathcal{P}_L \Xi, M \rangle| \leq 2\sqrt{(4c+8)} \sqrt{\frac{t}{nm}} \delta \sqrt{\varphi(k(\delta)+1)} + 2\sqrt{2}\delta \max_{v \in V} \|P_L e_v\| \frac{t}{n}.$$

Alternatively, if $k(\delta) \geq m$, we use (3.22) to get

$$|\langle \mathcal{P}_L \Xi, M \rangle| \leq 4\delta \sqrt{\frac{rt}{nm}} + 2\sqrt{2}\delta \max_{v \in V} \|P_L e_v\| \frac{t}{n}.$$

□

It follows from Lemma 3 that, for all $\delta > 0$, the following bound holds with probability at least $1 - e^{-t}$

$$\begin{aligned}
\sup_{\|M\|_2 \leq \delta, \|W^{1/2}M\|_2 \leq 1} |\langle \mathcal{P}_L \Xi, M \rangle| &\leq \\
2\sqrt{(4c+8)} \sqrt{\frac{t}{nm}} \delta \sqrt{\varphi(k(\delta)+1)} + 2\sqrt{2}\delta \max_{v \in V} \|P_L e_v\| \frac{t}{n} &
\end{aligned} \tag{3.24}$$

(recall that $\varphi(k) = r$ for $k \geq m$, so, the second bound of the lemma can be included in the first bound). Moreover, the bound can be easily made uniform in $\delta \in [\delta_-, \delta_+]$ for arbitrary $\delta_- < \delta_+$. To this end, take $\delta_j := \delta_+ 2^{-j}$, $j = 0, 1, \dots, \lceil \log_2(\delta_+/\delta_-) \rceil + 1$ and use (3.24) for each $\delta = \delta_j$ with $\bar{t} := t + \log(\lceil \log_2(\delta_+/\delta_-) \rceil + 2)$ instead of t . An application of the union bound and monotonicity of the left hand side and the right hand side of (3.24) with respect to δ then implies that with probability at least $1 - e^{-t}$ for all $\delta \in [\delta_-, \delta_+]$

$$\sup_{\|M\|_2 \leq \delta, \|W^{1/2}M\|_2 \leq 1} |\langle \mathcal{P}_L \Xi, M \rangle| \leq \quad (3.25)$$

$$C \sqrt{\frac{\bar{t}}{nm}} \delta \sqrt{\varphi(k(\delta) + 1)} + 4\sqrt{2}\delta \max_{v \in V} \|P_L e_v\| \frac{\bar{t}}{n}.$$

where $C > 0$ is a constant depending only on c . Indeed, by the union bound, (3.24) holds with probability at least

$$1 - (\lceil \log_2(\delta_+/\delta_-) \rceil + 2)e^{-\bar{t}} = 1 - e^{-t}$$

for all $\delta = \delta_j$, $j = 0, \dots, \lceil \log_2(\delta_+/\delta_-) \rceil + 1$. Therefore, for all $j = 0, \dots, \lceil \log_2(\delta_+/\delta_-) \rceil + 1$ and all $\delta \in (\delta_{j+1}, \delta_j]$

$$\sup_{\|M\|_2 \leq \delta, \|W^{1/2}M\|_2 \leq 1} |\langle \mathcal{P}_L \Xi, M \rangle| \leq \quad (3.26)$$

$$2\sqrt{(4c+8)} \sqrt{\frac{\bar{t}}{nm}} \delta_j \sqrt{\varphi(k(\delta_j) + 1)} + 2\sqrt{2}\delta_j \max_{v \in V} \|P_L e_v\| \frac{\bar{t}}{n}$$

(by monotonicity of the left hand side). Note that $k(\delta_j) \leq k(\delta) \leq k(\delta_{j+1})$. We can now use the fact that $\frac{\varphi(k)}{\lambda_k} = \frac{\varphi(k)}{k} \frac{k}{\lambda_k}$ is a nonincreasing function and the condition $\lambda_{k+1}/\lambda_k \leq c$ to show that

$$\begin{aligned} \sqrt{\frac{\bar{t}}{nm}} \delta_j \sqrt{\varphi(k(\delta_j) + 1)} &\leq 2\sqrt{\frac{\bar{t}}{nm}} \delta_{j+1} \sqrt{\varphi(k(\delta_{j+1}) + 1)} \leq \\ 2\sqrt{\frac{\bar{t}}{nm}} \sqrt{\frac{\varphi(k(\delta_{j+1}) + 1)}{\lambda_{k(\delta_{j+1})}}} &\leq 2\sqrt{c} \sqrt{\frac{\bar{t}}{nm}} \sqrt{\frac{\varphi(k(\delta_{j+1}) + 1)}{\lambda_{k(\delta_{j+1})+1}}} \\ 2\sqrt{c} \sqrt{\frac{\bar{t}}{nm}} \sqrt{\frac{\varphi(k(\delta) + 1)}{\lambda_{k(\delta)+1}}} &\leq 2\sqrt{c} \sqrt{\frac{\bar{t}}{nm}} \delta \sqrt{\varphi(k(\delta) + 1)}. \end{aligned}$$

This and bound (3.26) imply that

$$\sup_{\|M\|_2 \leq \delta, \|W^{1/2}M\|_2 \leq 1} |\langle \mathcal{P}_L \Xi, M \rangle| \leq \quad (3.27)$$

$$4\sqrt{c(4c+8)} \sqrt{\frac{\bar{t}}{nm}} \delta \sqrt{\varphi(k(\delta) + 1)} + 4\sqrt{2}\delta \max_{v \in V} \|P_L e_v\| \frac{\bar{t}}{n},$$

which proves bound (3.25).

Set δ as

$$\delta := \frac{\|\hat{S} - S_*\|_2}{\|W^{1/2}(\hat{S} - S_*)\|_2} = \frac{\|\hat{S} - S_*\|_{L_2(\Pi^2)}}{\|W^{1/2}(\hat{S} - S_*)\|_{L_2(\Pi^2)}}$$

and assume for now that $\delta \in [\delta_-, \delta_+]$. For a particular choice of $M := \frac{\hat{S} - S_*}{\|W^{1/2}(\hat{S} - S_*)\|_2}$, we get from (3.25) that

$$|\langle \mathcal{P}_L \Xi, \hat{S} - S_* \rangle| \leq C \sqrt{\frac{\bar{t}}{nm}} \|\hat{S} - S_*\|_2 \sqrt{\varphi(k(\delta) + 1)} + 4\sqrt{2} \max_{v \in V} \|P_L e_v\| \frac{\bar{t}}{n} \|\hat{S} - S_*\|_2. \quad (3.28)$$

Suppose now that $\delta^2 \geq \bar{\varepsilon}$. Since, under assumptions of the theorem, $\bar{\varepsilon} \in (\lambda_{s+1}^{-1}, \lambda_s^{-1}]$, this implies that $k(\delta) \leq k(\sqrt{\bar{\varepsilon}}) = s$ and

$$\begin{aligned} |\langle \mathcal{P}_L \Xi, \hat{S} - S_* \rangle| &\leq C \sqrt{\frac{\bar{t}}{nm}} \|\hat{S} - S_*\|_2 \sqrt{\varphi(s + 1)} + 4\sqrt{2} \max_{v \in V} \|P_L e_v\| \frac{\bar{t}}{n} \|\hat{S} - S_*\|_2 = \\ &C \sqrt{\frac{m\bar{t}}{n}} \|\hat{S} - S_*\|_{L_2(\Pi^2)} \sqrt{\varphi(s + 1)} + 4\sqrt{2} \max_{v \in V} \|P_L e_v\| \frac{m\bar{t}}{n} \|\hat{S} - S_*\|_{L_2(\Pi)} \leq \\ &2C^2 \frac{\varphi(s + 1)m\bar{t}}{n} + 64 \max_{v \in V} \|P_L e_v\|^2 \left(\frac{m\bar{t}}{n}\right)^2 + \frac{1}{4} \|\hat{S} - S_*\|_{L_2(\Pi^2)}^2. \end{aligned} \quad (3.29)$$

In the case when $\delta^2 < \bar{\varepsilon}$, we have $k(\delta) \geq k(\sqrt{\bar{\varepsilon}}) = s$. In this case, we again use the fact that $\frac{\varphi(k)}{\lambda_k}$ is a nonincreasing function and the condition $\lambda_{k+1}/\lambda_k \leq c$ to show that

$$\begin{aligned} \sqrt{\frac{\bar{t}}{nm}} \|\hat{S} - S_*\|_2 \sqrt{\varphi(k(\delta) + 1)} &= \sqrt{\frac{m\bar{t}}{n}} \|W^{1/2}(\hat{S} - S_*)\|_{L_2(\Pi^2)} \sqrt{\delta^2 \varphi(k(\delta) + 1)} \leq \\ &\sqrt{\frac{m\bar{t}}{n}} \|W^{1/2}(\hat{S} - S_*)\|_{L_2(\Pi^2)} \sqrt{\frac{\varphi(k(\delta) + 1)}{\lambda_{k(\delta)}}} \leq \sqrt{c} \sqrt{\frac{m\bar{t}}{n}} \|W^{1/2}(\hat{S} - S_*)\|_{L_2(\Pi^2)} \sqrt{\frac{\varphi(k(\delta) + 1)}{\lambda_{k(\delta)+1}}} \leq \\ &\sqrt{c} \sqrt{\frac{m\bar{t}}{n}} \|W^{1/2}(\hat{S} - S_*)\|_{L_2(\Pi^2)} \sqrt{\frac{\varphi(s + 1)}{\lambda_{s+1}}} \leq \sqrt{c} \sqrt{\frac{m\bar{t}}{n}} \sqrt{\bar{\varepsilon}} \|W^{1/2}(\hat{S} - S_*)\|_{L_2(\Pi^2)} \sqrt{\varphi(s + 1)}. \end{aligned}$$

This allows us to deduce from (3.28) that

$$\begin{aligned} |\langle \mathcal{P}_L \Xi, \hat{S} - S_* \rangle| &\leq \\ &\sqrt{c} C \sqrt{\frac{m\bar{t}}{n}} \sqrt{\bar{\varepsilon}} \|W^{1/2}(\hat{S} - S_*)\|_{L_2(\Pi^2)} \sqrt{\varphi(s + 1)} + 4\sqrt{2} \max_{v \in V} \|P_L e_v\| \frac{m\bar{t}}{n} \|\hat{S} - S_*\|_{L_2(\Pi)} \leq \\ &cC^2 \frac{\varphi(s + 1)m\bar{t}}{n} + \frac{1}{4} \bar{\varepsilon} \|W^{1/2}(\hat{S} - S_*)\|_{L_2(\Pi^2)}^2 + 32 \max_{v \in V} \|P_L e_v\|^2 \left(\frac{m\bar{t}}{n}\right)^2 + \frac{1}{4} \|\hat{S} - S_*\|_{L_2(\Pi^2)}^2. \end{aligned} \quad (3.30)$$

It follows from bounds (3.29) and (3.30) that with probability at least $1 - e^{-t}$,

$$\begin{aligned} |\langle \mathcal{P}_L \Xi, \hat{S} - S_* \rangle| &\leq (2 \vee c) C^2 \frac{\varphi(s + 1)m\bar{t}}{n} + 64 \max_{v \in V} \|P_L e_v\|^2 \left(\frac{m\bar{t}}{n}\right)^2 + \\ &\frac{1}{4} \|\hat{S} - S_*\|_{L_2(\Pi^2)}^2 + \frac{1}{4} \bar{\varepsilon} \|W^{1/2}(\hat{S} - S_*)\|_{L_2(\Pi^2)}^2, \end{aligned} \quad (3.31)$$

provided that

$$\delta = \frac{\|\hat{S} - S_*\|_2}{\|W^{1/2}(\hat{S} - S_*)\|_2} = \frac{\|\hat{S} - S_*\|_{L_2(\Pi^2)}}{\|W^{1/2}(\hat{S} - S_*)\|_{L_2(\Pi^2)}} \in [\delta_-, \delta_+]. \quad (3.32)$$

It remains now to substitute bounds (3.8), (3.10), (3.12) and (3.31) in bound (3.3) to get that with some constants $C > 0, C_1 > 0$ depending only on c and with probability at least $1 - 2e^{-t}$

$$\|\hat{S} - S_*\|_{L_2(\Pi^2)}^2 \leq C \frac{\varphi(s+1)m(\bar{t} + t_m)}{n} + \bar{\varepsilon} \|W^{1/2}S_*\|_{L_2(\Pi^2)}^2 + C_1 \max_{v \in V} \|P_L e_v\|^2 \left(\frac{m\bar{t}}{n}\right)^2, \quad (3.33)$$

where $t_m := t + \log(2m)$.

We still have to choose the values of δ_-, δ_+ and to handle the case when

$$\delta = \frac{\|\hat{S} - S_*\|_2}{\|W^{1/2}(\hat{S} - S_*)\|_2} = \frac{\|\hat{S} - S_*\|_{L_2(\Pi^2)}}{\|W^{1/2}(\hat{S} - S_*)\|_{L_2(\Pi^2)}} \notin [\delta_-, \delta_+]. \quad (3.34)$$

First note that, since the largest eigenvalue of W is λ_m and it is bounded from above by m^ζ , we have

$$\|W^{1/2}(\hat{S} - S_*)\|_2 \leq \sqrt{\lambda_m} \|\hat{S} - S_*\|_2 \leq m^{\zeta/2} \|\hat{S} - S_*\|_2.$$

Thus, $\delta \geq m^{-\zeta/2}$. Next note that

$$\|W^{1/2}S_*\|_{L_2(\Pi^2)}^2 \leq m^{-2}m^\zeta \|S_*\|_2^2 \leq m^\zeta,$$

where we also took into account that the absolute values of the entries of S_* are bounded by 1. It now follows from (3.14) that, under the assumption $\frac{2mt_m}{n} \leq 1$,

$$\begin{aligned} \|\hat{S} - S_*\|_{L_2(\Pi^2)}^2 &\leq \frac{3}{2}rm^2\varepsilon^2 + 2\bar{\varepsilon}m^\zeta \leq \\ 24rm^2 \frac{t + \log(2m)}{nm} + 2 \frac{m^\zeta}{\lambda_s} &\leq 12m + 2m^{2\zeta} \leq 14m^{2\zeta}, \end{aligned}$$

which holds with probability at least $1 - e^{-t}$. Therefore, as soon as $\|W^{1/2}(\hat{S} - S_*)\|_{L_2(\Pi^2)} \geq m^{-\zeta}$, we have $\delta \leq 4n^\zeta m^\zeta$.

We will now take $\delta_- := m^{-\zeta/2}, \delta_+ := 4n^\zeta m^\zeta$. Then, the only case when (3.34) can possibly hold is if $\|W^{1/2}(\hat{S} - S_*)\|_{L_2(\Pi^2)} \leq m^{-\zeta}$. In this case, we can set

$$\delta := n^\zeta \|\hat{S} - S_*\|_{L_2(\Pi^2)} \in [\delta_-, \delta_+]$$

and follow the proof of bound (3.31) replacing throughout the argument $\|W^{1/2}(\hat{S} - S_*)\|_{L_2(\Pi^2)}$ with $n^{-\zeta}$. This yields

$$\begin{aligned} |\langle \mathcal{P}_L \Xi, \hat{S} - S_* \rangle| &\leq \\ (2 \vee c) C^2 \frac{\varphi(s+1)m\bar{t}}{n} + 64 \max_{v \in V} \|P_L e_v\|^2 \left(\frac{m\bar{t}}{n}\right)^2 + \frac{1}{4} \|\hat{S} - S_*\|_{L_2(\Pi^2)}^2 + \frac{1}{4} \bar{\varepsilon} n^{-2\zeta}. \end{aligned} \quad (3.35)$$

Bound (3.35) can be now used instead of (3.31) to prove that

$$\|\hat{S} - S_*\|_{L_2(\Pi^2)}^2 \leq C \frac{\varphi(s+1)m(\bar{t} + t_m)}{n} + \bar{\varepsilon} \|W^{1/2} S_*\|_{L_2(\Pi^2)}^2 + C_1 \max_{v \in V} \|P_L e_v\|^2 \left(\frac{m\bar{t}}{n}\right)^2 + \bar{\varepsilon} n^{-2\zeta} \quad (3.36)$$

with some constants $C, C_1 > 0$ depending only on c .

Clearly, we can assume that $C_1 \geq 1$ and $\bar{t} \geq 1$. Since $m \leq n^2$ (recall that we even assumed that $mt_{n,m} \leq 1$), $\zeta \geq 1$, $\max_{v \in V} \|P_L e_v\|^2 \geq \frac{r}{m}^4$ and $\bar{\varepsilon} \leq \lambda_{k_0}^{-1} \leq m^\zeta$, it is easy to check that

$$C_1 \max_{v \in V} \|P_L e_v\|^2 \left(\frac{m\bar{t}}{n}\right)^2 \geq \frac{m}{n^2} \geq \frac{m^\zeta}{n^{2\zeta}} \geq \bar{\varepsilon} n^{-2\zeta}.$$

Thus, the last term of bound (3.36) can be dropped (with a proper adjustment of constant C_1).

Note also that with our choice of δ_-, δ_+

$$\bar{t} = t + \log(\log_2(\delta_+/\delta_- + 2)) \leq t + \log(\log_2(4n^\zeta m^{(3/2)\zeta}) + 2)$$

and $\bar{t} + t_m \leq 2t_{n,m}$. It is now easy to conclude that, with some constants C, C_1 depending only on c and with probability at least $1 - 3e^{-t}$

$$\|\hat{S} - S_*\|_{L_2(\Pi^2)}^2 \leq C \frac{\varphi(s+1)mt_{n,m}}{n} + \bar{\varepsilon} \|W^{1/2} S_*\|_{L_2(\Pi^2)}^2 + C_1 \max_{v \in V} \|P_L e_v\|^2 \left(\frac{m\bar{t}}{n}\right)^2. \quad (3.37)$$

The probability bound $1 - 3e^{-t}$ can be rewritten as $1 - e^{-t}$ by changing the value of constants C, C_1 . Also, by changing the notation $s+1 \mapsto s$, bound (3.37) yields (2.2). This completes the proof of the theorem. \square

References

- [1] Ahlswede, R. and Winter, A. (2002) Strong converse for identification via quantum channels. *IEEE Transactions on Information Theory*, 48, 3, pp. 569–679.

⁴Recall that $r = \|P_L\|_2^2 = \sum_{v \in V} \|P_L e_v\|^2$.

- [2] Aubin, J.-P. and Ekeland, I. (1984) Applied Nonlinear Analysis. J. Wiley&Sons, New York.
- [3] Candes, E. and Recht, B. (2009) Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6), 717–772.
- [4] Candes, E. and Tao, T. (2010) The power of convex relaxation: Near-optimal matrix completion. *IEEE Transactions on Information Theory*, 56, 2053–2080.
- [5] Candes, E. and Plan, Y. (2011) Tight Oracle Bounds for Low-Rank Matrix Recovery from a Minimal Number of Random Measurements. *IEEE Transactions on Information Theory*, 57(4), 2342–2359.
- [6] Gross, D. (2011) Recovering Low-Rank Matrices From Few Coefficients in Any Basis. *IEEE Transactions on Information Theory*, 57, 3, 1548–1566.
- [7] Koltchinskii, V. (2011a) Von Neumann Entropy Penalization and Low Rank Matrix Estimation. *Annals of Statistics*, 39, 6, 2936–2973.
- [8] Koltchinskii, V. (2011b) Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems, *Ecole d’ete de Probabilités de Saint-Flour 2008*, Lecture Notes in Mathematics, Springer.
- [9] Koltchinskii, V. (2011c) A remark on low rank matrix recovery and noncommutative Bernstein type inequalities. Preprint.
- [10] Koltchinskii, V., Lounici, K. and Tsybakov, A. (2011) Nuclear norm penalization and optimal rates for noisy matrix completion. *Annals of Statistics*, 39, 5, 2302–2329.
- [11] Negahban, S. and Wainwright, M.J. (2010) Restricted strong convexity and weighted matrix completion with noise. Preprint.
- [12] Recht, B., Fazel, M. and Parrilo, P. (2010) Guaranteed minimum rank solutions of matrix equations via nuclear norm minimization. *SIAM Review*, 52, 3, 471–501.
- [13] Rohde, A. and Tsybakov, A. (2011) Estimation of high-dimensional low rank matrices. *Annals of Statistics*, 39, 2, 887–930.
- [14] Tropp, J.A. (2010) User-friendly tail bounds for sums of random matrices. *Foundations of Computational Mathematics*, to appear.